SISMID Exercises

George Dewey,

2024-06-11

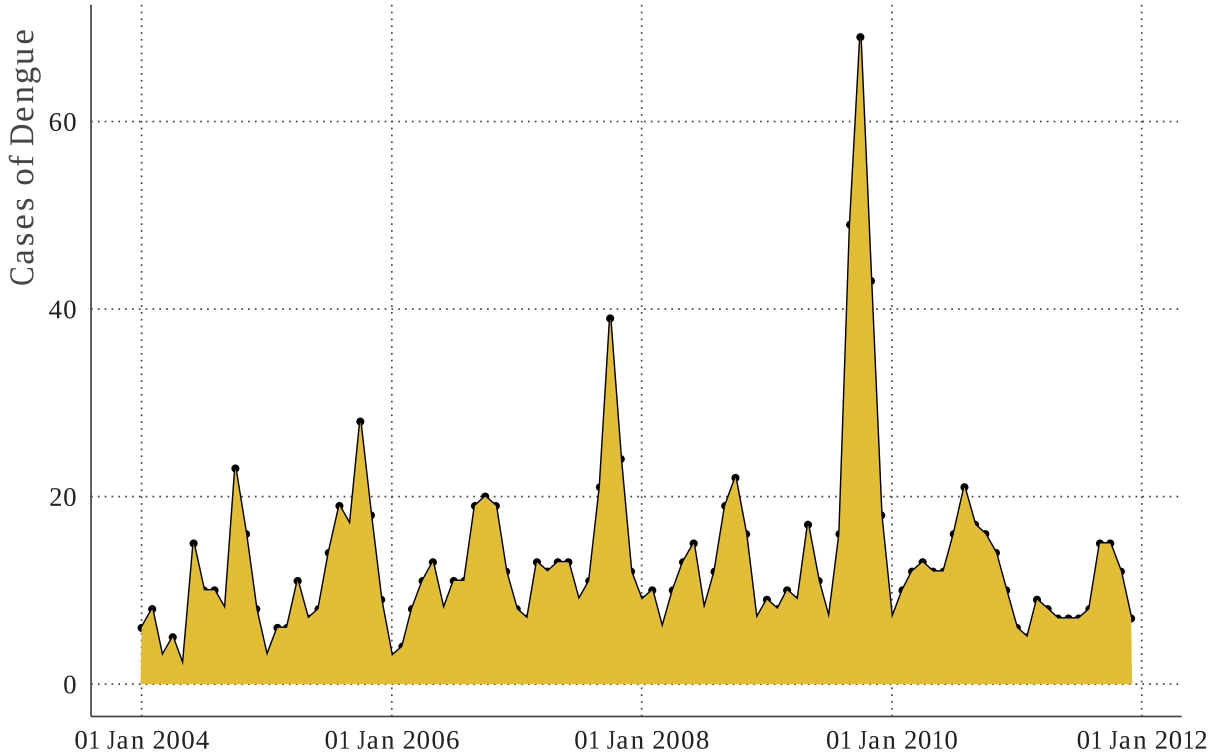
Exercise 1

```
data_e1 = read_csv('/Users/gdewey/Documents/Projects/SISMID23/data/MX_Dengue_trends.csv', show_col_type

data_e1 = data_e1 %>% rename(dengue_true = `Dengue CDC`,
                             date = Date)
```

a) Plot the number of cases of Dengue as a function of time.

```
ggplot(aes(x = date, y = dengue), data = data_e1) +
  geom_line() +
  geom_point(size = 0.75) +
  geom_area(fill = firaPalette()[5]) +
  xlab('') +
  ylab('Cases of Dengue') +
  scale_x_date(date_labels = '%d %b %Y') +
  theme_fira()
```



b) For the training period 2004-2006 (36 months), find the best line that explains the number of cases of Dengue as a function of the number of searches of the term "dengue". You should do this by solving the least squares problem, and you should obtain the value of the y-intercept and the slope.

```r
data_e1_lm = data_e1 %>% slice(1:36)

m1 = lm(dengue_true ~ dengue, data = data_e1_lm)
coefs_lm = coef(m1)

coefs_lm['(Intercept)']
```

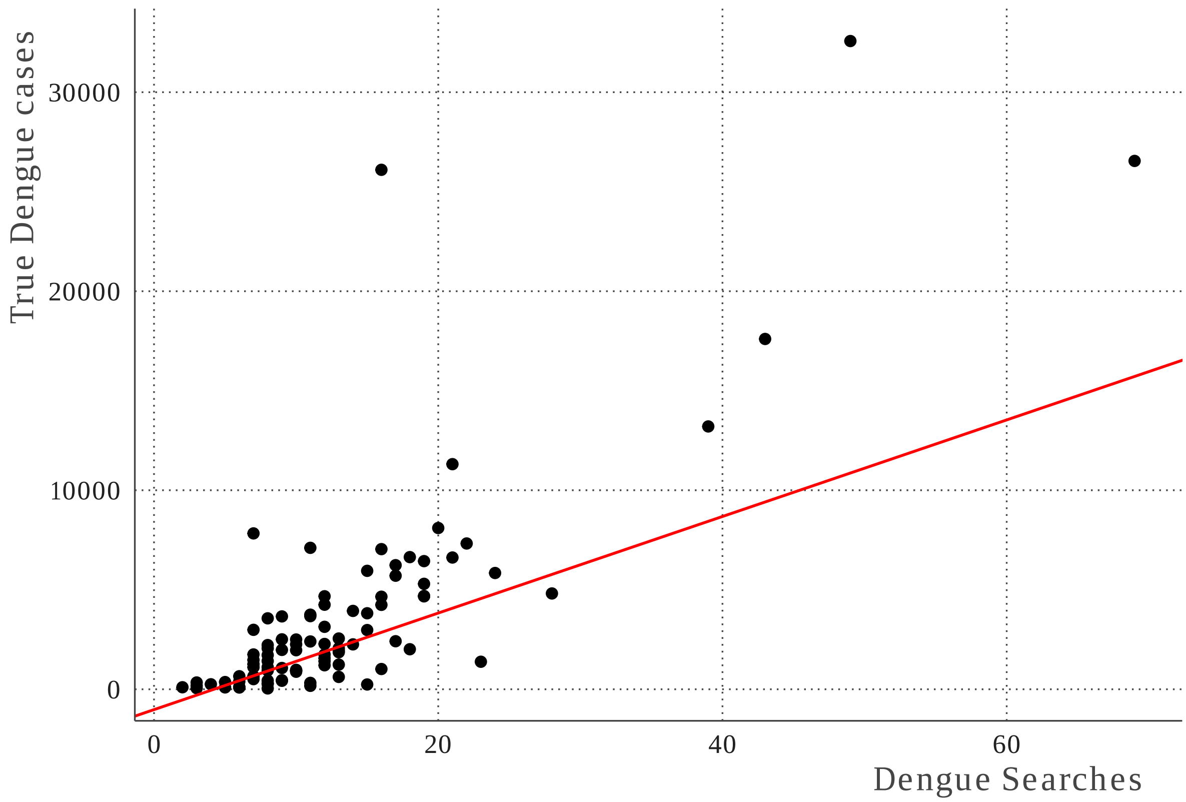
```
## (Intercept)
##     -1023.9
```

```r
coefs_lm['dengue']
```

```
##   dengue
## 242.6151
```

c) Use the equation of the line you obtained in (b) and plot the number of cases as a function of the number of searches of the term "dengue", predicted by your method during the training period. Compare your results to the plot in (a) for such time period.

```
ggplot(aes(x = dengue, y = dengue_true), data = data_e1) +
  geom_point() +
  geom_abline(intercept = -1023.9, slope = 242.6151, color = 'red') +
  theme_minimal() +
  xlab('Dengue Searches') +
  ylab('True Dengue cases') +
  theme_fira()
```

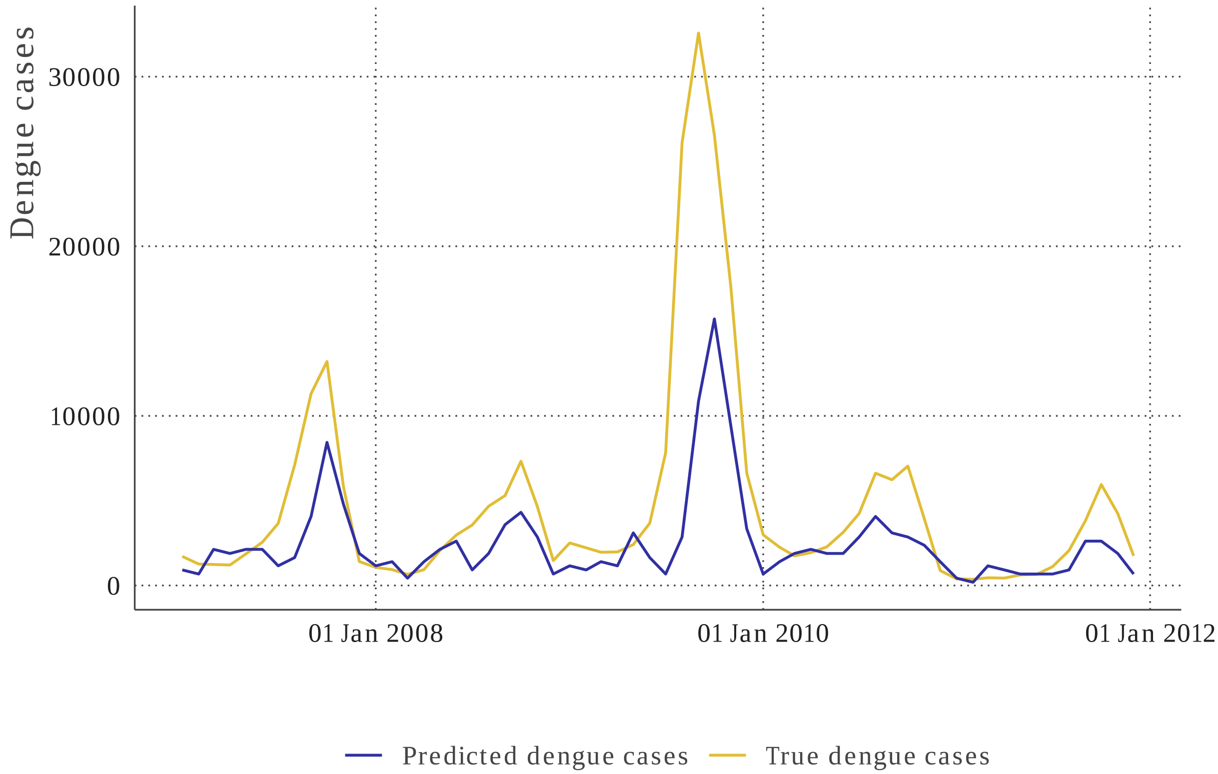


The red line represents the prediction generated by the linear regression model; increasing the search count increases the magnitude of under-prediction.

d) For the prediction or validation period 2007-2011, use the equation of the line you obtained in (b) to predict the number of the dengue cases as a function of the number of searches of the term "dengue" from 2007-2011. Plot your predictions and compare them to the actual number of cases.

```
data_e1_pred = data_e1 %>% slice(37:nrow(data_e1))
data_e1_pred$predicted_vals = predict(m1, data.frame(dengue = data_e1_pred$dengue))
```

```
ggplot(aes(x = date), data = data_e1_pred) +
  geom_line(aes(y = dengue_true, color = 'True dengue cases')) +
  geom_line(aes(y = predicted_vals, color = 'Predicted dengue cases')) +
  xlab('') +
  ylab('Dengue cases') +
  scale_color_fira(name = '') +
  scale_x_date(date_labels = '%d %b %Y') +
  theme_fira() +
  theme(legend.position = 'bottom')
```



e) Discuss your results. Could you improve this modeling approach? If so, how?

In general, the predicted values are lower than the true values, suggesting that more data sources are needed to achieve a more accurate prediction. Aggregation of other data sources (like the counts of other search terms in our dataset) could improve the accuracy of predictions.